Machine Learning

AIML CZG565

M8 : Bayesian Learning

BITS Pilani
Pilani Campus

Course Faculty of M.Tech Cluster
BITS – CSIS - WILP

Machine Learning

innovate achieve lead

Disclaimer and Acknowledgement

•These content of modules & context under topics are planned by the course owner Dr. Sugata, with grateful acknowledgement to many others who made their course materials freely available online

•We here by acknowledge all the contributors for their material and inputs.

•We have provided source information wherever necessary

•Students are requested to refer to the textbook w.r.t detailed content of the presentation deck shared over canvas

•We have reduced the slides from canvas and modified the content flow to suit the requirements of the course and for ease of class presentation

Slide Source / Preparation / Review:
From BITS Pilani WILP**:** Prof.Sugata, Prof.Chetana, , Prof.Monali, Prof.Rajavadhana

External: CS109 and CS229 Stanford lecture notes, Dr.Andrew NG and many others who made their course materials freely available online

Course Plan

M1	Introduction & Mathematical Preliminaries
M2	Machine Learning Workflow
M3	Linear Models for Regression
M4	Linear Models for Classification
M5	Decision Tree
M6	Instance Based Learning
M7	Support Vector Machine
M8	Bayesian Learning
M9	Ensemble Learning
M10	Unsupervised Learning
M11	Machine Learning Model Evaluation/Comparison

Bayesian Learning
Parameter Estimation

- Where does the **cost** come from? - Logistic regression
- Why least-squares **cost** function, be a reasonable choice? – Linear regression

BITS Pilani

Logistic Regression

sigmoid function $y = \frac{1}{1+e^{-z}}$ where $z = w^T x + b$

↓

O/P prob $\in [0, 1]$

Drawback

least squares

* get stuck in local minima
* prob. doesn't match

(i) Maximum Likelihood Estimation ✓

Linear Regression

$y = ax + b$

$\hat{y}$ → predicted

$E = (y - \hat{y})^2$

error

(1) cost fn. is convex ✓ unique minimum

(2) differentiable

$|y - \hat{y}| = 0.1$

$(y - \hat{y})^2 = (0.1)^2 = 0.01$

$\frac{1}{m}|y - \hat{y}|$

Why we need a general rule?

- In real-world problems, we often **don't have such a simple formula** like "7 out of 10".
  We might have:
- Continuous data (e.g., heights, prices, sensor readings)
- Complex models (e.g., regression, neural networks)
- Multiple unknown parameters (e.g., slope, intercept, variance)
- In all these cases, we can't "just count".
  We need a **systematic, principled way** to estimate parameters — that's what MLE gives us.

So — MLE is **the universal mathematical rule** for parameter estimation.

Distribution

R.V : Input → ℝ

f(x):

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

Numeric

Mileage (in kmpl)	*Car Price (in cr)*
Neutral	High
Less	Low
Neutral	Medium
More	Low

Non-numeric

Mileage (in kmpl)	*Car Price (in cr)*
9.8	High
9.12	Low
9.5	High
10	Low

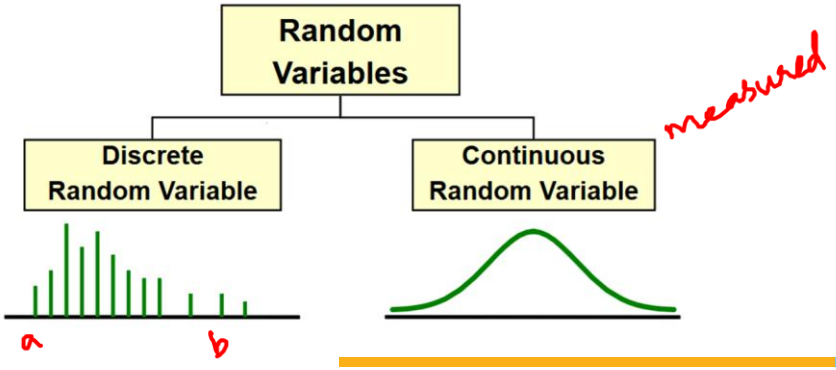
mixed → data

$P(a \leq X \leq b)$ = count

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

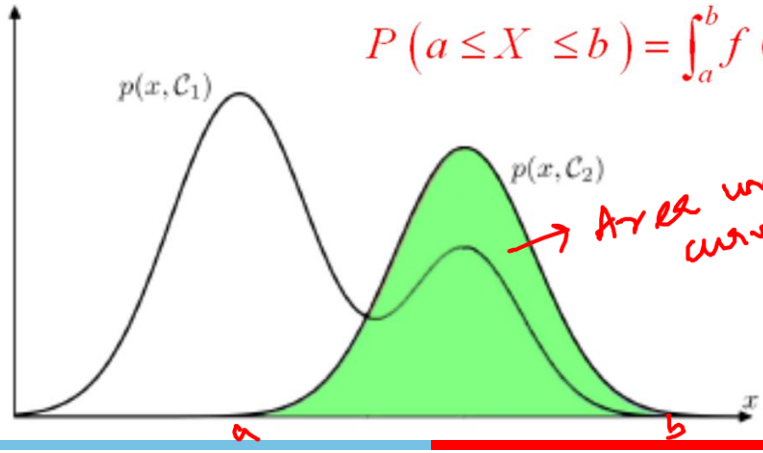
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Represents a possible numerical value from a random event

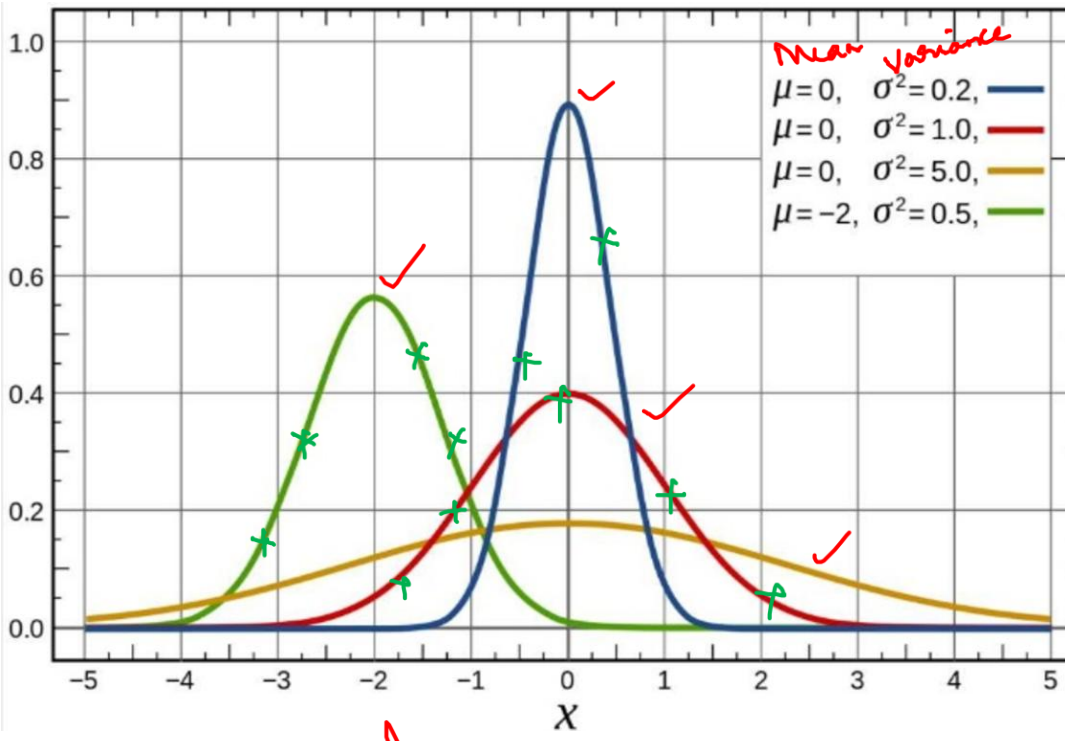


measured

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$



Parameter Estimation



Distribution	Parameters
Bernoulli(p)	$\theta = p$
Poisson(λ)	$\theta = \lambda$
Uniform(a,b)	$\theta = (a, b)$
Normal(μ, σ^2)	$\theta = (\mu, \sigma^2)$
Y = mX + b	$\theta = (m, b)$

θ is the parameter of a distribution.

θ can be a vector of parameters

Distribution = model + parameter θ

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Find $\boldsymbol{\theta} = (\mathbf{Mean}, \mathbf{SD})$ from the data $\mathbf{X_i}$

ie., $(x_i, P(x_i))$

y → line

o/p of end result

shape/structure

Distribution = model + parameter θ

$y \in 0/1$

Linear Regression

Model : Normal distribution

Parameters: (m, b, σ^2)

$E = \Sigma(y - \hat{y})^2$

Sum of small random effects will tend to follow normal distribution based on (Central limit theorem)

CLT

Large sample $n \geq 30$

Logistic Regression

Model : Bernoulli distribution

(P) → success

Parameters: (w, b)

Maximum Likelihood Estimation

- Think of MLE as the **bridge between probability theory and data**:

It's like reverse-engineering nature:

- Nature gives you data (e.g., coin flips, car prices, etc.)
- You assume a generative model (Binomial, Gaussian, etc.)

MLE says: "Okay, which version of this model best fits what nature gave me?"

Parameter in ML

Mileage (in kmpl)	Car Price (in cr)
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

$$\text{CarPrice} = 8.5 + 0.5\ \text{Mileage} - 1.5\ \text{Mileage}^2$$

$\theta_0 + \theta_1\ \text{Mileage} + \theta_2\ \text{Mileage}^2$

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \rightarrow$ Linear regression

$$\text{Parameters} : (\theta_0, \theta_1, \theta_2)$$

Mileage (in kmpl)	Car Price (in cr)
9.8	High 1
9.12	Low 0
9.5	High 1
10	Low 0

classification

Logistic Regression

$$\text{CarPrice} = \frac{1}{1+e^{-8.5 + 0.5\ \text{Mileage} - 1.5\ \text{Mileage}^2}}$$

$$= \frac{1}{1+e^{-z}}$$

Parameter Estimation in ML

IID → identically Independent

Mileage (in kmpl)	Car Price (in cr)
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

CarPrice = 8.5 + 0.5 Mileage – 1.5 Mileage^2

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \varepsilon_i$

$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$

$\varepsilon^{(i)} \sim N(0, \sigma^2)$

Parameters : $(\theta_0, \theta_1, \theta_2)$

Find $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ from the data $\mathbf{X_i}$

ie., (Mileage$_i$, CarPrice$_i$)

Assumption: Data are__

- IID samples: $X_1 \ldots X_n$ where all X_i are independent and have the same distribution.
- Either same PMF (discrete) or same PDF (continuous)
- f (X | θ)
  Likelihood of different values of X depends on the values of our parameters θ
  f( ) is either PDF or PMF

Parameter Estimation

Maximum Likelihood Estimation (MLE)

select that parameters θ that make the observed data the most likely

$f(X_1, X_2 \ldots .X_n \mid \boldsymbol{\theta})$

Maximum A Posteriori (MAP)

choose the parameters θ that is the most likely, given the data

$f(\boldsymbol{\theta} \mid X_1, X_2 \ldots .X_n)$

PMF → discrete R.V $\quad \sum p(x) = 1$

X	0	1	2	3
$P(X)$	0	$1/2$	$1/4$	$1/2$

PDF → continuous R.V $\quad x = 0.43 \quad x = 0.01$

$$\int_0^1 f(x)\,dx = 1$$

$$f(x) = \begin{cases} 1/2 & 0 \le x \le 0.5 \\ 3/4 & 0.5 \le x \le 1 \end{cases}$$

MLE and MAP

Intuitive Picture

Imagine a **family of curves** (distributions), each labelled by a different θ.

Your observed data is a **point on the x-axis**.

MLE: Choose the curve that passes **highest at your data point** (makes the data most probable).

MAP: Choose the curve that is **both high at your data point and high in prior plausibility**.

Key intuition

- **MLE**: "Which curve explains the data best?"
- **MAP**: "Which curve explains the data well *and* was believable to begin with?"

Why this matters?

MAP avoids overfitting because it:

- Penalizes unlikely/complex parameter values
- Incorporates prior knowledge

One-line summary

- MLE = best fit to data
  MAP = best tradeoff between data fit and prior belief

Intuition of Bayes Theorem

$P(Y|X)$ → already happened

MAP : f (θ | X_1 , X_2 .....X_n)

MLE: f (X_1 , X_2 .....X_n | θ)

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h) = prior probability of hypothesis h

P(D) = prior probability of training data D

P(h|D) = probability of h given D

P(D|h) = probability of D given h

Total probability

$L(\theta)$ probability of data given parameter θ

After seeing data, posterior belief of θ

$$\underbrace{P(\theta|\text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data}|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{P(\text{data})}$$

Before seeing data, prior belief of θ e.g. what is distribution over parameters θ

Total probability

Maximum Likelihood Estimation (MLE)

$L(\theta) \rightarrow$ Likelihood function

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve for θ

MLE – Linear Regression Model

Given the noise $\varepsilon^{(i)}$ obeys a Normal distribution each $y^{(i)}$ must also obey a Normal distribution around the true target value

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

CarPrice = 8.5 + 0.5 Mileage – 1.5 Mileage^2

$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$

$\varepsilon^{(i)} \sim N(0, \sigma^2)$

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right).$$

Parameters : $(\theta_0, \theta_1, \theta_2)$

Find $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ from the data $\mathbf{X_i}$

ie., (Mileage$_i$, CarPrice$_i$)

$\mathbf{y^{(i)} \mid x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)}$

Maximum Likelihood Estimation (MLE)

select that parameters θ that make the observed data the most likely

f (X_1 , X_2 .....X_n | $\boldsymbol{\theta}$)

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

MLE – Linear Regression Model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$

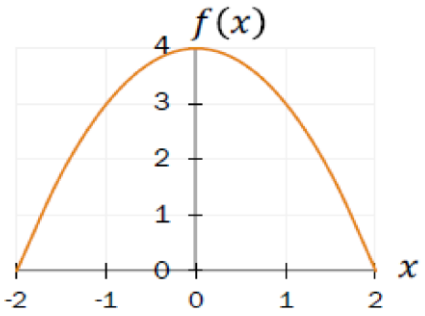
$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta)$$

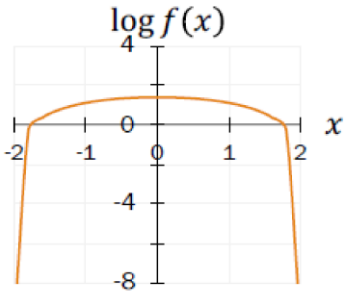
2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) \qquad LL(\theta) = \log L(\theta)$$

3. Solve





Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

MLE answers the question: For which parameter value does the observed data have the largest probability?

MLE – Linear Regression Model

$\log e^{x} = x$

$\log(ab) = \log a + \log b$

$\log a^{n} = n \log a$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve for θ

$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, L(\theta) \qquad LL(\theta) = \log L(\theta)$$

$$= \log L(\theta)$$

$\log L(\theta)$

$$= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

•Θ:A parameter (or set of parameters) of a model.
•**L(θ)**: A *likelihood function*—it measures how well the parameter θ explains the observed data.
•**argmaxθ**: "The value of θ that maximizes …"

θ ^ is the value of θ that makes the likelihood function $L(\theta)$as large as possible.

Problem: Estimate a coin's bias

Let:

$$\theta = P(heads)$$

You flip the coin 5 times and observe:

H, H, T, H, T → 3 heads, 2 tails

Success p failure (1−p)

1. Write the likelihood function

The likelihood $L(\theta)$ is the probability of observing this exact sequence given θ.

For independent flips: $L(\theta)=\theta^3(1-\theta)^2$

$$L(\theta) = \theta^3 (1-\theta)^2$$

2. Try some values of θ

Let's see how the likelihood behaves:

If $\theta = 0.5$:

$$L = (0.5)^3(0.5)^2 = 0.5^5 = 0.03125$$

If $\theta = 0.6$:

$$L = (0.6)^3(0.4)^2 = 0.03456$$

If $\theta = 0.7$:

$$L = (0.7)^3(0.3)^2 = 0.03087$$

👉 The likelihood is highest around **0.6**

3. Solve it exactly (using calculus)

Maximize:

$$L(\theta) = \theta^3(1-\theta)^2$$

It's easier to maximize the **log-likelihood**:

$$\log L(\theta) = 3\log\theta + 2\log(1-\theta)$$

Take derivative:

$$\frac{d}{d\theta} = \frac{3}{\theta} - \frac{2}{1-\theta}$$

Set to zero:

$$\frac{3}{\theta} = \frac{2}{1-\theta}$$

$\frac{d}{dx}[\log x] = \frac{1}{x}$

$\frac{d}{d\theta}[\log\theta] = \frac{1}{\theta}$

$\frac{3}{\theta} - \frac{2}{1-\theta} = 0$

$3(1-\theta) - 2\theta = 0$

$3 - 3\theta - 2\theta = 0$

$3 - 5\theta = 0$

Solve:

$$3(1-\theta) = 2\theta \Rightarrow 3 - 3\theta = 2\theta \Rightarrow 3 = 5\theta \Rightarrow \hat{\theta} = \frac{3}{5} = 0.6$$

Final answer

$$\hat{\theta} = 0.6$$

MLE – Linear Regression Model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$

$$\log L(\theta) = n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

independent of θ

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial}{\partial\theta}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2\right]$$

$$\nabla_\theta L(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\theta^T\mathbf{x}^{(i)})\mathbf{x}^{(i)}$$

$\frac{d}{dx}(x^2) = 2x$

3. Solve

Set $\nabla_\theta L(\theta) = 0$

find the value 'θ'

$$\frac{2(y^{(i)}-\theta^T x^{(i)})(x^{(i)})}{2\sigma^2}$$

MLE – Linear Regression Model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$

$$\log L(\theta) = n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ $= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$

$$\nabla_\theta L(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\theta^T \mathbf{x}^{(i)})\mathbf{x}^{(i)}$$

3. Solve

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, L(\theta)$$

$$= \text{argmax}\ \left(-\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2\right)$$

$$= \text{argmin}\ \left(\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2\right)$$

expression

mean squared error

With probabilistic assumptions on the data, least-squares regression corresponds to finding the MLE of θ

MLE – Logistic Regression

$$y_i \mid x_i \sim \text{Bern}(\sigma(\mathbf{w}^\top x_i))$$

1. Determine formula for $LL(\theta)$ $\quad LL(\theta) = \log \prod_{i=1}^{n} P_\theta\left(y^{(i)}|x^{(i)}\right) = \sum_{i=1}^{n} \log P_\theta\left(y^{(i)}|x^{(i)}\right)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ $\quad p(y \mid x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$
3. Solve

Mileage (in kmpl)	*Car Price (in cr)*
9.8	High
9.12	Low
9.5	High
10	Low

Bernoulli MLE Estimation

$X_1, X_2, \ldots, X_n$ where $X_i \sim$ Ber(p). PMF of a Bernoulli $p^{X_i}(1-p)^{1-X_i}$

$$P_\theta(Y = 1|X = x) = h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}$$

$$P_\theta(Y = 0|X = x) = 1 - h_\theta(x) = \frac{e^{-\theta^\top x}}{1 + e^{-\theta^\top x}}$$

MLE answers the question: For which parameter value does the observed data have the largest probability?

MLE – Logistic Regression

$\frac{d}{dx}(\log x) = \frac{1}{x}\,dx$ $\quad \log a^b = b \log a$

$$y_i \mid x_i \sim \text{Bern}(\sigma(\mathbf{w}^\intercal x_i))$$

1. Determine formula for $LL(\theta)$

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} | X = \mathbf{x}^{(i)})$$

$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

$h_\theta(x)$

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

1. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} y \log \sigma(\theta^T \mathbf{x}) + \frac{\partial}{\partial \theta_j} (1 - y) \log[1 - \sigma(\theta^T \mathbf{x}]$$

$$= \left[\frac{y}{\sigma(\theta^T \mathbf{x})} - \frac{1 - y}{1 - \sigma(\theta^T \mathbf{x})}\right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T \mathbf{x})$$

$$= \left[\frac{y}{\sigma(\theta^T \mathbf{x})} - \frac{1 - y}{1 - \sigma(\theta^T \mathbf{x})}\right] \sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$

$$= \left[\frac{y - \sigma(\theta^T \mathbf{x})}{\sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]}\right] \sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$

$$= [y - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$

$\frac{\partial LL(\theta)}{\partial \theta_j} = 0$

2. Solve

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta)$$

MLE answers the question: For which parameter value does the observed data have the largest probability?

MLE – Logistic Regression

$$y_i \mid x_i \sim \text{Bern}(\sigma(\mathbf{w}^\intercal x_i))$$

1. Determine formula for $LL(\theta)$

$$L(\theta) = \prod_{i=1} P(Y = y^{(i)} | X = \mathbf{x}^{(i)})$$
$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

1. Differentiate $LL(\theta)$ w.r.t. (each) θ $\quad \dfrac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)})\right] x_j^{(i)}$

2. Solve for θ .

MLE answers the question: For which parameter value does the observed data have the biggest probability?

MLE-Summary

- Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, drawn from a distribution $f(X_i|\theta)$

- θ_{MLE} maximizes the likelihood of data, $L(\theta)$ and $LL(\theta)$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

Maximum A Posteriori (MAP) Analysis

1. Determine prior probability

$$\theta_{MAP} = \arg\max f(\theta|X_1, X_2, \ldots, X_n)$$

1. Find the posterior probability for every distinct prior

Brute Force MAP Hypothesis

$$h_{MAP} = \underset{h \in H}{argmax}\ P(h|D)$$

$$= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)}$$

$$= \underset{h \in H}{argmax}\ P(D|h)P(h)$$

1. Choose the posterior with highest h_{MAP} value

Maximum a Posteriori (MAP) Estimator of θ is the value of θ that maximizes the posterior distribution of θ.

Best hypothesis ≈ most probable hypothesis

Maximum A Posteriori Estimation (MAP)

1. Find the prior probability ✓

1. Derive the posterior probability ✓

1. Differentiate posterior w.r.t. (each) θ ✓

2. Solve

Maximum a Posteriori (MAP) Estimator of θ is the value of θ that maximizes the posterior distribution of θ.

Best hypothesis ≈ most probable hypothesis

Example :

1. Example on MAP algorithm:

Let X be continuous random variable with probability density function P(X) given by:

$$P(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Given another distribution $p(Y|X = x) = x(1-x)^{y-1}$ Find MAP estimate of X given Y=3

solution

h_{MAP} = P(X | Y=3)

= P(Y=3 | X) * P(X)

= $x(1-x)^{y-1} * 2x$

To find the parameter X , differentiate the function & equate to zero.

$$\frac{d(P(X|Y=3)}{dx} = 0$$

$$\frac{d(x(1-x)^2 * 2x)}{dx} = 0$$

$$\frac{d(2x^2-4x^3+2X^4)}{dx} = 0$$

$$4x-12x^2+8X^3 = 0$$

$x = \{0, 0.5, 1\}$

Example :

1. Example on MAP algorithm:

Let X be continuous random variable with probability density function P(X) given by:

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Given another distribution $p(Y|X = x) = x(1-x)^{y-1}$ Find MAP estimate of X given Y=3

h_{MAP} = P(X | Y=3)
= P(Y=3 | X) * P(X)
= $x(1-x)^{y-1} * 2x$

x = {0, 0.5, 1} ✓

P(X|Y=3) = {0, 0.125 ,0} → output/ final answer.

When $x = 0.5$

$h_{MAP} = 0.5(1-0.5)^2[2(0.5)]$

$= 0.5(0.5)^2$

ML setting

- Bayesian Analysis
  - start with some belief about the system, called a prior.
  - Then we obtain some data and use it to update our belief.
  - The outcome is called a posterior.
  - Should we obtain even more data, the old posterior becomes a new prior and the cycle repeats.
  - P(h | D) a posterior determines the class label
  - MLE and MAP are the same if the prior is uniform
  - This forms the basis for Naïve Bayes classifier More on this will be discussed in later part of this module

Bayes' Optimal Classifier

•The most probable classification of the new instance is obtained by combining **the predictions of all hypotheses, weighted by their posterior probabilities.**

•v_j from some set V, then the probability P(v_j I D) that the correct classification for the new instance is v_j is: $$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

•The optimal classification of the new instance is the value vj for which P(v_j I D) is maximum

Most Probable Classification of New Instances

- So far we've sought the most probable hypothesis given the data D (i.e., h_{MAP})
- **Given new instance x, what is its most probable classification?**
  - $h_{MAP}(x)$ is not the most probable classification!

    using the single best model, h_{MAP}, to make a prediction **ignores all the other hypotheses** in H, even if they have slightly lower, but still significant, posterior probabilities P(h|D).
  - What's most probable classification of x?

Consider:

Three possible hypotheses:

$P(h_1|D) = .4$, $P(h_2|D) = .3$, $P(h_3|D) = .3$

Given new instance x, classification given by above 3 hypotheses is

$h_1(x) = +$, $h_2(x) = -$, $h_3(x) = -$

$P(\oplus|h_1) = 1$ and $P(\ominus|h_1) = 0$

May be the classification for X is **+**

Example 1 : Bayes Optimal Classifier

Bayes optimal classification: $\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$

- Example:

$P(h_1|D) = .4,\ P(-|h_1) = 0,\ P(+|h_1) = 1$

$P(h_2|D) = .3,\ P(-|h_2) = 1,\ P(+|h_2) = 0$

$P(h_3|D) = .3,\ P(-|h_3) = 1,\ P(+|h_3) = 0$

therefore

and

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

A. Total Probability for Class '+'

We sum the weighted probabilities for all hypotheses that predict '+':

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D)$$

- For h_1: $P(+|h_1)P(h_1|D) = 1 \times 0.4 = 0.4$
- For h_2: $P(+|h_2)P(h_2|D) = 0 \times 0.3 = 0$.
- For h_3: $P(+|h_3)P(h_3|D) = 0 \times 0.3 = 0$

Total Probability for '+' is: $0.4 + 0 + 0 = \mathbf{0.4}$

B. Total Probability for Class '−

We sum the weighted probabilities for all hypotheses that predict '−':

$$\sum_{h_i \in H} P(-|h_i)P(h_i|D)$$

- For h_1: $P(-|h_1)P(h_1|D) = 0 \times 0.4 = 0$
- For h_2: $P(-|h_2)P(h_2|D) = 1 \times 0.3 = 0.3$
- For h_3: $P(-|h_3)P(h_3|D) = 1 \times 0.3 = 0.3$

Total Probability for '−' is: $0 + 0.3 + 0.3 = \mathbf{0.6}$

Gibbs Classifier

- Bayes optimal classifier provides best result, but can be expensive if many hypotheses.
- Gibbs algorithm:
  - Choose one hypothesis at random, according to posterior prob. Distribution over h , $P(h|D)$
  - Use this h to classify new instance
- Surprising fact: under certain conditions, the expected misclassification error for the Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier

$$E[error_{Gibbs}] \leq 2E[error_{BayesOptional}]$$

- Suppose correct, uniform prior distribution over H, then
  - Pick any hypothesis from *Version space*, with uniform probability
  - Its expected error no worse than twice Bayes optimal

Parameter Estimation in ML - Summary

Assumption: Data are__

- IID samples: $X_1 \ldots X_n$
  where all X_i are independent and have the same distribution.
- Either same PMF (discrete) or same PDF (continuous)
- f (X | θ)
  Likelihood of different values of X depends on the values of our parameters θ
  f( ) is either PDF or PMF

Maximum Likelihood Estimation (MLE)
select that parameters θ that make the observed data the most likely

Maximum A Posteriori (MAP)
choose the parameters θ that is the most likely, given the data

If the sample is large, MLE will yield an excellent estimator of θ
When no prior information is available, all hypothesis are equally likely i.e. p(hi) = p(hj)

Additional Problems in MLE

BITS Pilani

MLE – Discrete - PMF

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ. $\theta = 0.5$

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

MLE – Discrete - PMF

1. Determine formula for $LL(\theta)$

   L(ɵ) = P(X=3)*P(X=0)*P(X=2)*……..*P(X=1)

   $= ((1-\theta) / 3)^2 * (2\theta / 3)^2 * (2(1-\theta) / 3)^3 * (\theta / 3)^3$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

X	P(X)
3	(1-ɵ) / 3
0	2ɵ / 3
2	2(1-ɵ) / 3
1	ɵ/ 3
3	(1-ɵ) / 3
2	2(1-ɵ) / 3
1	ɵ/ 3
0	2ɵ / 3
2	2(1-ɵ) / 3
1	ɵ/ 3

MLE – Discrete - PMF

$$L(\theta) = P(X=3)*P(X=0)*P(X=2)*\ldots\ldots*P(X=1)$$

$$= ((1-\theta)/3)^2 * (2\theta/3)^2 * (2(1-\theta)/3)^3 * (\theta/3)^3$$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \log[((1-\theta)/3)^2 * (2\theta/3)^2 * (2(1-\theta)/3)^3 * (\theta/3)^3]$$

$$= \log((1-\theta)/3)^2 + \log(2\theta/3)^2 + \log(2(1-\theta)/3)^3 * \log(\theta/3)^3$$

$$= 2\log((1-\theta)/3) + 2\log(2\theta/3) + 3\log(2(1-\theta)/3) + 3\log(\theta/3)$$

$$=2(\log((1-\theta) - \log 3) + 2(\log(2\theta) - \log 3) + 3(\log(2(1-\theta)) - \log 3) + 3(\log(\theta) -$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\text{Gradient}(LL(\theta)) = -\frac{2}{(1-\theta)} + \frac{2}{\theta} - \frac{3}{(1-\theta)} + \frac{3}{\theta}$$

$$= \frac{-5\theta + 5 - 5\theta}{(1-\theta)\theta} = \frac{-10\theta + 5}{(1-\theta)\theta} = 0 \rightarrow \theta = 0.5$$

3. Solve

$\frac{d}{d\theta}[2\log 2\theta] = \frac{2}{2\theta} d\theta$

$\frac{d}{dx}[\log x] = \frac{1}{x} dx$

$\log(ab) = \log a + \log b$

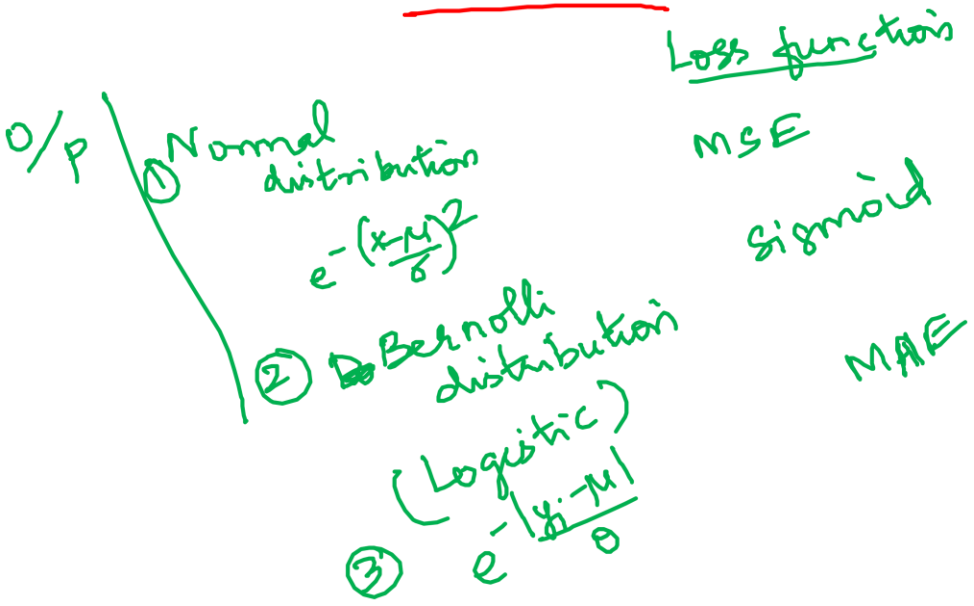
$\log a^x = x \log a$

Additional Exercise for Student's Practice

Consider inputs x_i which are real valued attributes and the outputs y_i which are real valued of the form $y_i = f(x_i) + e_i$, where $f(x_i)$ is the true function and e_i is a random variable representing laplacian noise with PDF given by

$$f(y_i/\theta) = \frac{1}{2\theta} * e^{\frac{-|y_i - \mu|}{\theta}}$$

Implementing a linear regression model of the form , $h(x_i) = \sum_{i=0}^{n} \theta_i x_i$ and $\mu = h_\theta(x_i)$ find the maximum likelihood estimator of è. Comment on the loss function.



Additional Exercise for Student's Practice

1. Determine formula for $LL(\theta)$

$$L(\theta) = L(\theta) = \frac{1}{2\theta} e^{\frac{-|y_1-\mu|}{\theta}} * \frac{1}{2\theta} e^{\frac{-|y_2-\mu|}{\theta}} * \frac{1}{2\theta} e^{\frac{-|y_3-\mu|}{\theta}} * \ldots\ldots * \frac{1}{2\theta} e^{\frac{-|y_n-\mu|}{\theta}}$$

$$= \left(\prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|y_i-\mu|}{\theta}}\right)$$

1. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$LL(\theta) = \ln\left(\prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|y_i-\mu|}{\theta}}\right) = -\ln(2\theta) \sum_{i=1}^{n} 1 + \sum_{i=1}^{n} \ln\left(e^{\frac{-|y_i-\mu|}{\theta}}\right)$$

$$= \underset{\theta}{argmax} - n \ln(2\theta) - \sum_{i=1}^{n} \frac{|y_i-\mu|}{\theta}$$

$$= \underset{\theta}{argmin}\ n \ln(2\theta) + \sum_{i=1}^{n} \frac{|y_i-\mu|}{\theta}$$

1. Solve

$$\text{Gradient}\ (\ LL(\theta)\ ) = \frac{n}{(\theta)} - \frac{\sum_{i=1}^{n} |y_i-\mu|}{(\theta^2)} = 0$$

$$\frac{n}{(\theta)} = \frac{\sum_{i=1}^{n} |y_i-\mu|}{(\theta^2)}$$

$$\theta = \frac{\sum_{i=1}^{n} |y_i-\mu|}{n}$$

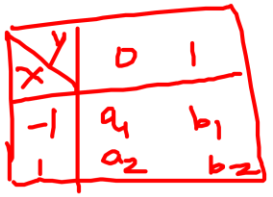
Instead of MSE, MAE is the maximum likelihood hypothesis. So MAE is appropriate for the loss function

Naïve Bayes Classifier

BITS Pilani

Probability Basics

- Prior, conditional and joint probability
  - Prior probability:
  - Conditional probability: $P(x|C)$
  - Joint probability: $P(x, y) = a_1$
  - Relationship: $y = f(x) = 2x$ $y = x^2$
  - Independence: $P(A)$ does affect the $P(B)$.



- Bayesian Rule

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})}$$

Prior

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Probability

Probabilistic Classification

- Establishing a probabilistic model for classification
  - Discriminative model
    $$P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$
  - Generative model
    $$P(\mathbf{X} \mid C) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$
- MAP classification rule
  - MAP: Maximum A Posterior
  - Assign x to c^* if $P(C = c^* \mid \mathbf{X} = \mathbf{x}) > P(C = c \mid \mathbf{X} = \mathbf{x}) \quad c \neq c^*, \; c = c_1, \cdots, c_L$
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert:
    $$P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})} \propto P(\mathbf{X} \mid C)P(C)$$

Naïve Bayes

Is it discriminative or generative model? $P(X, Y)$

supervised classification algorithm.

- Bayes classification

joint conditional probability of the features

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_n \mid C)P(C)$$

Difficulty: learning the joint probability — computationally intractable.

- Naïve Bayes classification
  - Making the assumption that all input attributes are independent

$$\begin{aligned} P(X_1, X_2, \cdots, X_n \mid C) &= P(X_1 \mid X_2, \cdots, X_n; C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2 \mid C)\cdots P(X_n \mid C) \end{aligned}$$

individual probabilities

Maximum A Posterior

MAP classification rule

$$[P(x_1 \mid c^*)\cdots P(x_n \mid c^*)]P(c^*) > [P(x_1 \mid c)\cdots P(x_n \mid c)]P(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Classifying emails as spam (C) or not spam ($\bar{C}$)

Let X_1 = "contains the word free"

X_2 = "contains a link"

X_3 = "contains the word win"

label
spam
spam
not spam

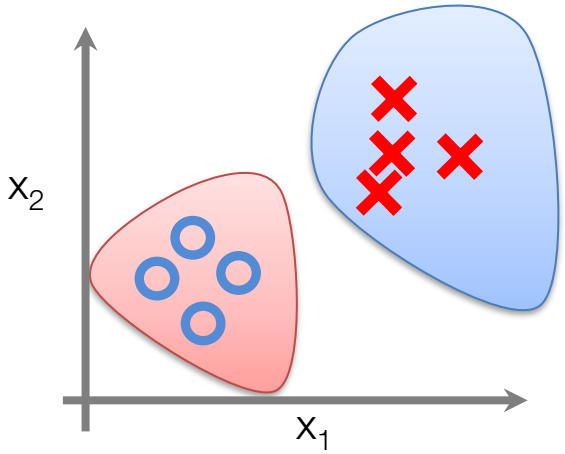
Then $P(X_1, X_2, X_3 \mid C = \text{spam}) = 2/3$

$P(X_1, X_2, X_3 \mid \bar{C} = \text{not spam}) = 1/3$

} joint conditional probability of all features given class C or $\bar{C}$.

Decision Theory: Generative Models

Generative



Likelihood — Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability — Predictor Prior Probability

Total law of prob

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

$P(Y \mid x_1, x_2 \ldots x_n) =$

$$P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y)P(Y)}{P(X_1 X_2 \ldots X_d)}$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes ??

Test

Known as generative models, because by sampling from them it is possible to generate synthetic data points in the input space.

Eg., Gaussians, **Naïve Bayes**, Mixtures of multinomials **, Mixtures of Gaussians**, Bayesian networks

Generative models for classification

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

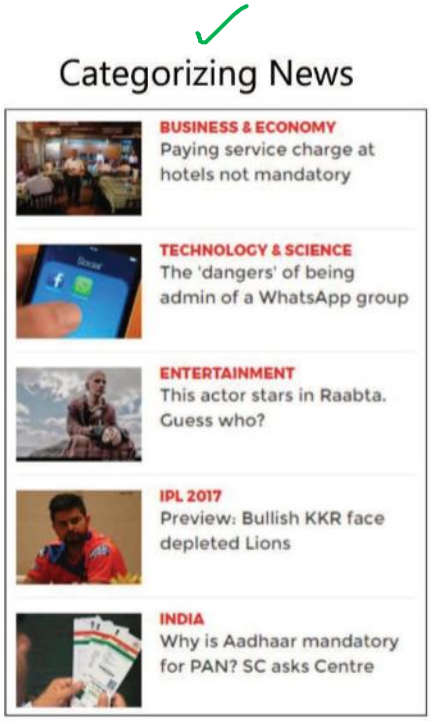
- For binary classification the denominator is given by

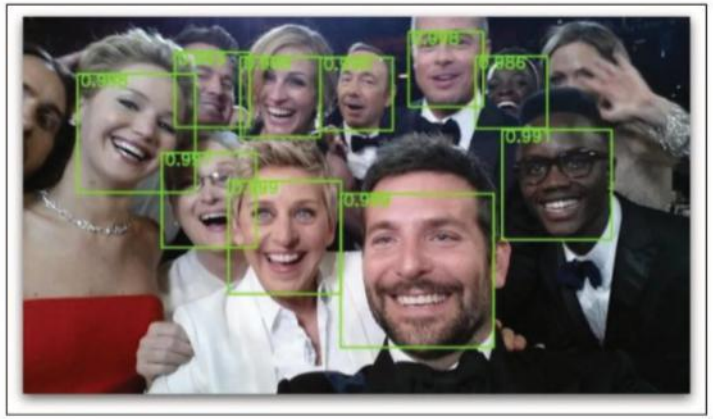
$$p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$$

- if were calculating p(y|x) in order to make a prediction, then we don't actually need to calculate the denominator, since

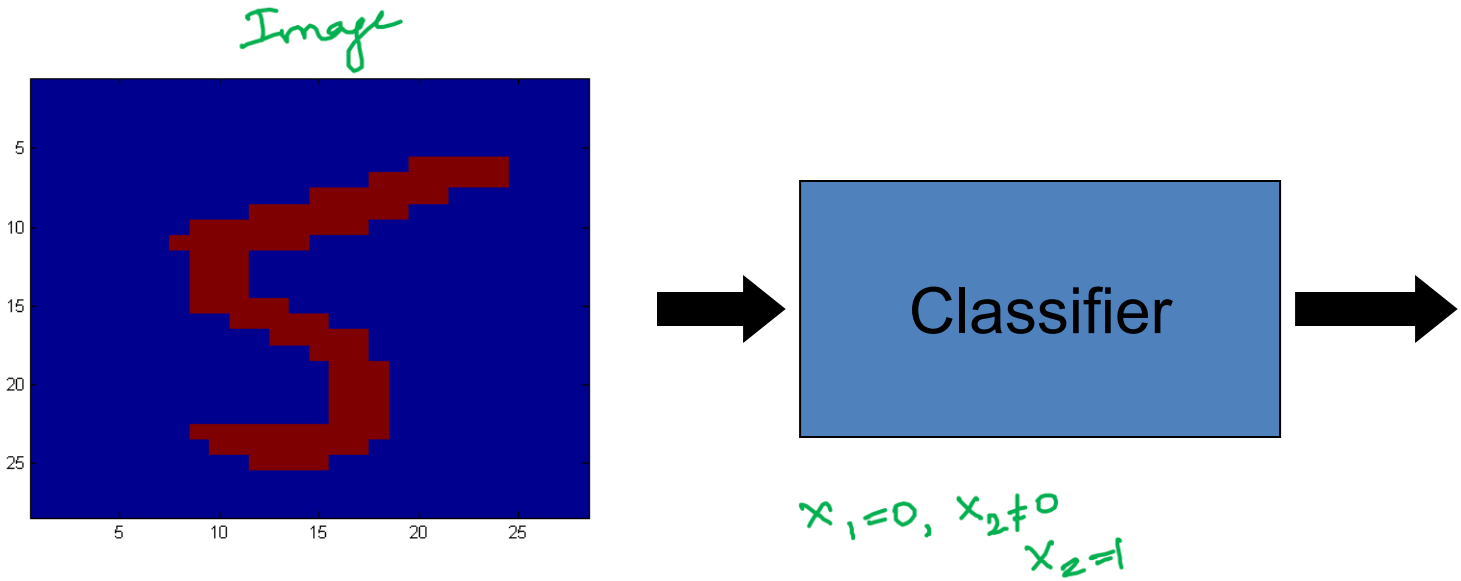
$$\begin{aligned} \arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y). \end{aligned}$$

Naïve Bayes Classifier - Applications


Categorizing News
BUSINESS & ECONOMY
Paying service charge at hotels not mandatory
TECHNOLOGY & SCIENCE
The 'dangers' of being admin of a WhatsApp group
ENTERTAINMENT
This actor stars in Raabta. Guess who?
IPL 2017
Preview: Bullish KKR face depleted Lions
INDIA
Why is Aadhaar mandatory for PAN? SC asks Centre



Example: Digit Recognition



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

Digit Recognition

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

$$P(Y = 5|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|Y = 5)P(Y = 5)}{P(X_1, \ldots, X_n|Y = 5)P(Y = 5) + P(X_1, \ldots, X_n|Y = 6)P(Y = 6)}$$

→ total prob.

$$P(Y = 6|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|Y = 6)P(Y = 6)}{P(X_1, \ldots, X_n|Y = 5)P(Y = 5) + P(X_1, \ldots, X_n|Y = 6)P(Y = 6)}$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

Naïve Bayes conditional Independence assumption

- Naïve Bayes assumes X_i are conditionally independent given Y

$$P(X_1|X_2, Y) = P(X_1|Y)$$

X_1 = tossing a coin
X_2 = rolling a dice

X_1	X_2	Y
H	1	Win
H	2	Lose
T	1	Win
T	3	loss

- **Assumption**:

$$P(X_1, \cdots, X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$$

$= P(X_1|Y) \cdot P(X_2|Y) \ldots P(X_n|Y)$ $\quad P(X_1, X_2|Y)$

i.e., X_i and X_j are conditionally independent given Y for $i \neq j$

Slide credit: Tom Mitchell

Naïve Bayes classifier: Prediction

Goal of learning P(Y|X) where X = <X_1,…, X_n>

- Bayes rule:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \cdots, X_n | Y = y_j)}$$

- Assume conditional independence among X_i's:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \Pi_i P(X_i | Y = y_j)}$$

- Classify New Instance(x) : Pick the most probable (MAP) Y for

$$X_{new} = < X_1, \ldots, X_n >$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Prior

Likelihood

Slide credit: Tom Mitchell

Example: Play Tennis

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(X|Y) \sim \text{Multinom}(\pi, n) \rightarrow \text{Multinomial NB } (X_i - \text{multinomial})$$

$$P(Y) \sim \text{Ber}(p$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Example: Play Tennis – Learning Phase

P(Sunny | Yes) = #(Sunny ∧ Yes) / #(Yes)

Look up tables

Maximum likelihood estimates (MLE's):

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}} \qquad \hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

class label

Sky	Play=Yes	Play=No
Sunny	3/3	2/4
Rainy	0/3	2/4

Number of items in dataset D for which $Y=y_k$

Wind	Play=Yes	Play=No
Strong	2/3	3/4
Breeze	1/3	1/4

Total o/p → 7

Yes = 3

No = 4

AirTemp	Play=Yes	Play=No
Hot	0/3	1/4
Warm	3/3	1/4
Cold	0/3	2/4

Forecast	Play=Yes	Play=No
Same	2/3	2/4
Change	1/3	2/4

Humidity	Play=Yes	Play=No
High	1/3	3/4
Normal	2/3	1/4

P(Yes) = 3/7

P(No) = 4/7

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport? (output)
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes / No ??

$P(Y|X) = P(X|Y)\,P(Y)$

Example: Play Tennis - Testing Phase

MAP rule

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)
= P(X | Enjoy=Yes). P(Enjoy=Yes)
= P(X | Enjoy=Yes). (3/7)
= P(Sunny | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Strong | Enjoy=Yes). P(Change | Enjoy=Yes).(3/7)
= (3/3) . (3/3) . (2/3) . (2/3) . (1/3) . (3/7)
=0.0635

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

P(Enjoy=Yes | X) > P(Enjoy=No | X) ⇒ EnjoySport = **Yes**

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

Training set

Test set

Yes

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)
= P(X | Enjoy=No). P(Enjoy=No)
= P(X | Enjoy=No). (4/7)
= P(Sunny | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Strong | Enjoy=No). P(Change | Enjoy=No). (4/7)
= (2/4) . (1/4) . (1/4) . (3/4) . (2/4) . (4/7)
= 0.006696

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples) for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

Example: Play Tennis

$\frac{1}{2} = \frac{2}{4} = \frac{4}{8}$

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (0+1/3) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

Sky	Enjoy Sport?
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Rainy	????

AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Warm	Normal	Strong	Same	Yes
Warm	High	Strong	Same	No
Cold	High	Strong	Change	No
Warm	Normal	Breeze	Same	Yes
Hot	Normal	Breeze	Same	No
Cold	High	Strong	Change	No
Warm	High	Strong	Change	Yes
Warm	Normal	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2+1/4) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

Laplace Smoothing

Smoothing

If one of the conditional probabilities is zero, then the entire expression becomes zero

- Technique for smoothing categorical data.
- A small-sample correction, or **pseudo-count**, will be incorporated in every probability estimate.
- No probability will be zero.

Smoothing

Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + d}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

Bayesian approach

d: number of distinct values of an attribute

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class *c*

p: prior probability of the class

m: constant called the **equivalent sample size**, which determines how heavily to weight p relative to the observed data

Example: Play Tennis

$P(Enjoy=Yes \mid X) = P(X \mid Enjoy=Yes).\ P(Enjoy=Yes) / P(X)$

$= P(X \mid Enjoy=Yes).\ P(Enjoy=Yes)$

$= P(X \mid Enjoy=Yes).\ (3/7)$

$= P(Rainy \mid Enjoy=Yes).\ P(Warm \mid Enjoy=Yes).\ P(Normal \mid Enjoy=Yes).\ P(Breeze \mid Enjoy=Yes).\ P(Same \mid Enjoy=Yes).\ (3/7)$

$= (0+1/3+2)\ .\ (3+1/3+3)\ .\ (2+1/3+2)\ .\ (1+1/3+2)\ .\ (2+1/3+2)\ .\ (3+1/7+2)$

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

$P(Enjoy=No \mid X) = P(X \mid Enjoy=No).\ P(Enjoy=No) / P(X)$

$= P(X \mid Enjoy=No).\ P(Enjoy=No)$

$= P(X \mid Enjoy=No).\ (4/7)$

$= P(Rainy \mid Enjoy=No).\ P(Warm \mid Enjoy=No).\ P(Normal \mid Enjoy=No).\ P(Breeze \mid Enjoy=No).\ P(Same \mid Enjoy=No).\ (4/7)$

$= (2+1/4+2)\ .\ (1+1/4+3)\ .\ (1+1/4+2)\ .\ (1+1/4+2)\ .\ (2+1/4+2)\ .\ (4+1/7+2)$

Example: Play Tennis

Test data 2

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

P(Enjoy=Yes | X) > P(Enjoy=No | X) ⇒ EnjoySport = **Yes**

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). Enjoy=No). (4/7)

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

Yes

Naïve Bayes: Continuous Features

- Continuous-valued Input Attributes
  - Numberless values for an attribute
  - Conditional probability modeled with the normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

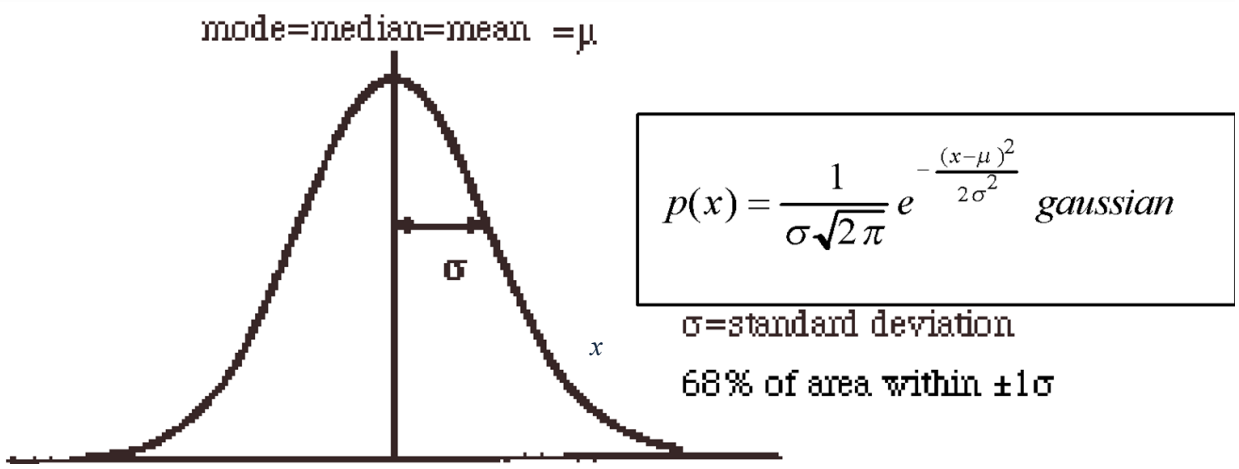
μ_{ji} : mean (avearage) of attribute values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of attribute values X_j of examples for which $C = c_i$

  - Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_n)$, $C = c_1, \cdots, c_L$
    Output: $n \times L$ normal distributions and $P(C = c_i)$ $i = 1, \cdots, L$
  - Test Phase: for $\mathbf{X}' = (X_1', \cdots, X_n')$
    - Calculate conditional probabilities with all the normal distributions
    - Apply the MAP rule to make a decision

The Gaussian Probability Distribution

- Gaussian Naïve Bayes

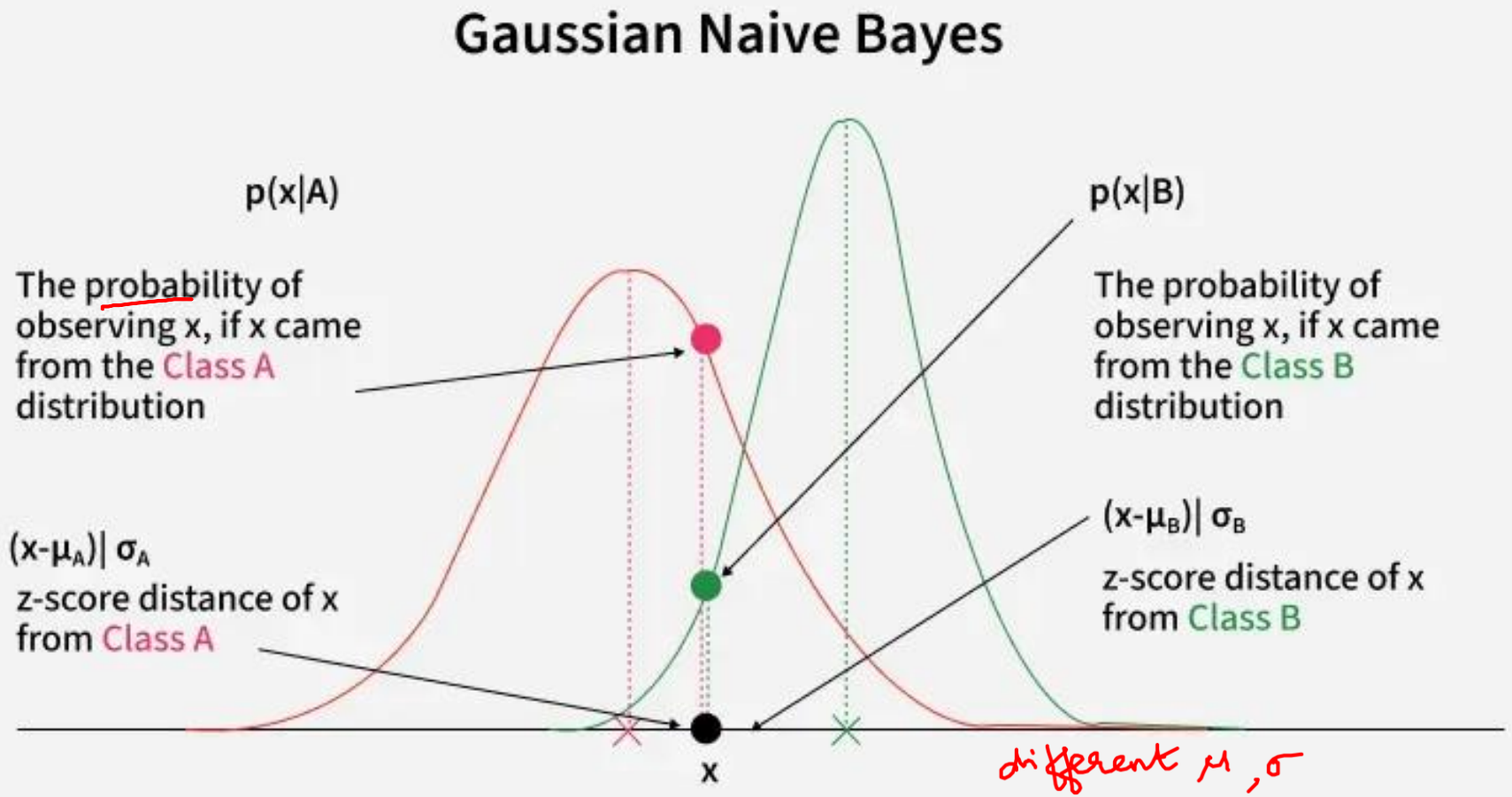


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad gaussian$$

σ=standard deviation

68% of area within ±1σ

Continuous Features : learning


Gaussian Naive Bayes
p(x|A)
The probability of observing x, if x came from the Class A distribution
p(x|B)
The probability of observing x, if x came from the Class B distribution
$(x-\mu_A)|\ \sigma_A$
z-score distance of x from Class A
$(x-\mu_B)|\ \sigma_B$
z-score distance of x from Class B
x
different μ, σ

Ref: https://www.geeksforgeeks.org/machine-learning/gaussian-naive-bayes/

Continuous Features : learning

To understand how Gaussian Naive Bayes works here's a simple binary classification problem using one feature: petal length.

Petal Length (cm)	Class Label
1.4	0 (Iris-setosa)
1.3	0 (Iris-setosa)
1.5	0 (Iris-setosa)
4.5	1 (Iris-versicolor)
4.7	1 (Iris-versicolor)
4.6	1 (Iris-versicolor)

1. Separate by Class

- $Class0: [1.4, 1.3, 1.5]$
- $Class1: [4.5, 4.7, 4.6]$

2. **.Calculate Mean and Variance**

For class 0:

- $\mu_0 = \frac{1.4+1.3+1.5}{3} = 1.4$
- $\sigma_0^2 = \frac{\left(1.4-1.4\right)^2+\left(1.3-1.4\right)^2+\left(1.5-1.4\right)^2}{3} = 0.0067$

For class 1:

- $\mu_1 = \frac{4.5+4.7+4.6}{3} = 4.6$
- $\sigma_1^2 = \frac{\left(4.5-4.6\right)^2+\left(4.7-4.6\right)^2+\left(4.6-4.6\right)^2}{3} = 0.0067$

$\sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n}$

$P(C=0) = 3/6 = 1/2$

$P(C=1) = 3/6 = 1/2$

Continuous Features : learning

Let the new instance be 1.6

3. Gaussian Likelihood

The Gaussian PDF is:

$$P(x, \mid \mu\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

·1e−2σ2(x−μ)2

For $x = 1.6$: New instance

$\mu = 1.4$ $\sigma^2 = 0.0067$

Class 0

- $P(1.6 \mid C = 0) \approx \frac{1}{\sqrt{2\pi \cdot 0.0067}} \cdot e^{-\frac{(1.6-1.4)^2}{2 \cdot 0.0067}} \approx 0.247$
- ··1e−2·0.0067(1.6−1.4)2≈0.247

Class 1

- $P(1.6 \mid C = 1) \approx \frac{1}{\sqrt{2\pi \cdot 0.0067}} \cdot e^{-\frac{(1.6-4.6)^2}{2 \cdot 0.0067}} \approx 0$ ✓
- ··1e−2·0.0067(1.6−4.6)2≈0 ✓

$P(X|C=0)$ (0.5)

$P(X|C=1)$ (0.5)

$P(C|X) = P(X|C)\,P(C)$

4. Multiply by Class Priors

Assume equal priors:

$$P(C = 0) = P(C = 1) = 0.5$$

Then:

- $P(C = 0 \mid x) \propto 0.247 \cdot 0.5 = 0.1235$ ✓
- $P(C = 1 \mid x) \propto 0 \cdot 0.5 = 0$

.5Prediction

Since $P(C = 0 \mid x) > P(C = 1 \mid x)$,

$x = 1.6 \in$ Iris-setosa

Predicted Class: $C = 0$ *(Iris−setosa)*

Example: Play Tennis

What should we do when one of the prob. is zero in Gaussian Naive Bayes?

$P(Enjoy=Yes \mid X) = P(X \mid Enjoy=Yes).\ P(Enjoy=Yes) / P(X)$

$= P(X \mid Enjoy=Yes).\ P(Enjoy=Yes)$

$= P(X \mid Enjoy=Yes).\ (3/7)$

$= P(Rainy \mid Enjoy=Yes).\ P(Warm \mid Enjoy=Yes).\ P(60 \mid Enjoy=Yes).\ P(Breeze \mid Enjoy=Yes).\ P(Same \mid Enjoy=Yes).(3/7)$

$= (1/3)\ .\ (3/3)\ .\ 0.15*10^{-95}\ .\ (1/3)\ .\ (2/3)\ .\ (3/7)$

Gaussian Naive Bayes

Humidity	Enjoy Sport?	AirTemp	Sky	Wind	Forecast	Enjoy Sport?
85	Yes	Warm	Sunny	Strong	Same	Yes
80	No	Warm	Sunny	Strong	Same	No
70	No	Cold	Rainy	Strong	Change	No
83	Yes	Warm	Rainy	Breeze	Same	Yes
90	No	Hot	Sunny	Breeze	Same	No
50	No	Cold	Rainy	Strong	Change	No
85	Yes	Warm	Sunny	Strong	Change	Yes
60	????	Warm	Rainy	Breeze	Same	????

$P(Enjoy=No \mid X) = P(X \mid Enjoy=No).\ P(Enjoy=No) / P(X)$

$= P(X \mid Enjoy=No).\ P(Enjoy=No)$

$= P(X \mid Enjoy=No).\ (4/7)$

$= P(Rainy \mid Enjoy=No).\ P(Warm \mid Enjoy=No).\ P(60 \mid Enjoy=No).\ P(Breeze \mid Enjoy=No).\ P(Same \mid Enjoy=No).\ (4/7)$

$= (2/4)\ .\ (1/4)\ .\ 0.02\ .\ (1/4)\ .\ (2/4)\ .\ (4/7)$

Additional Exercises – For Student's Practice

As a part of efforts to improve students' performance in the exams, you have been given the data showing number of study hours spent by students, their gender and their final results as pass or fail. Using this sample dataset, apply Naïve Bayes classification technique, to classify the below test case:

{No of study hours = 3.5, Gender="male"} either as "Pass", or "Fail".

Tip: $P(X|Y) \sim N(\mu, \sigma^2) \rightarrow$ GaussianNB (X_i – real valued)

No of study hours	Gender	Final result
4.5	Male	Pass
7	Female	Pass
2	Male	Fail
4	Female	Fail
2.5	Male	Fail
3	Female	Fail
8.3	Male	Fail
8	Female	Pass
9	Male	Pass

Additional Exercises – For Student's Practice

P(Pass | X) = P(X | Pass). P(Pass) / P(X)
= P(X | Pass). P(Pass)
= P(X | Pass). (0.44)
= P(Male | Pass). P(3.5 | Pass).(0.44)
= (2/4). 0.1056 . (0.44) = 0.0235

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

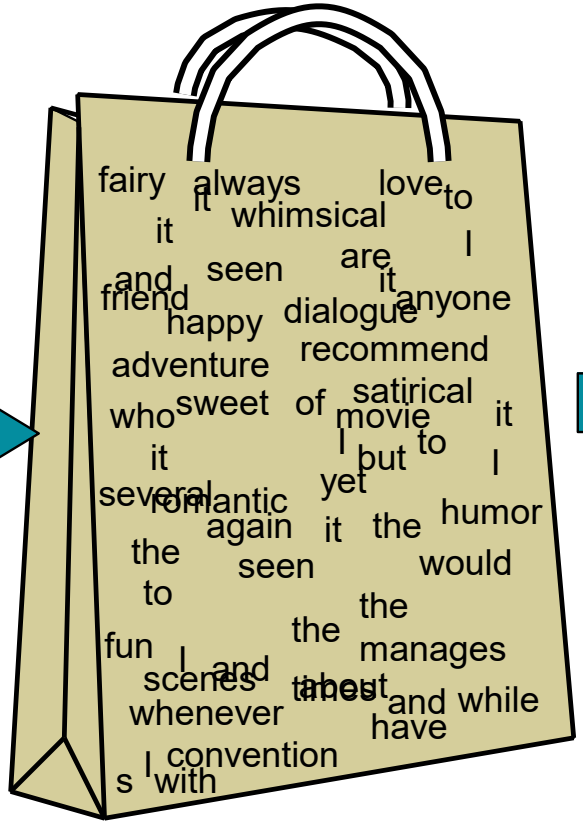
P(Fail | X) = P(X | Fail). P(Fail) / P(X)
= P(X | Fail). P(Fail)
= P(X | Fail). (0.56)
= P(Male | Fail). P(3.5 | Fail).(0.56)
= (3/5). 0.1846 . (0.56) = 0.0615

No of study hours	Gender	Final result
4.5	Male	Pass
7	Female	Pass
2	Male	Fail
4	Female	Fail
2.5	Male	Fail
3	Female	Fail
8.3	Male	Fail
8	Female	Pass
9	Male	Pass

Text Classification using Naive Bayes Classifier

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsic al	1
times	1
sweet	1
satirical	1
adventur e	1
genre	1
fairy	1
humor	1
have	1

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

Text	Tag
"A great game"	Sports ✓
"The election was over"	Not sports
"Very clean match"	Sports ✓
"A clean but forgettable game"	Sports ✓
"It was a close election"	Not sports

P(Sports=Yes) = 3/5
P(Sports=No) = 2/5
P(A | Sports=Yes) = 2/11
P(A | Sports=No) = 1/9

11 = total words in "Sports"

9 = total words in "not sports"

P(Sports=Yes | X) = P(X | Sports=Yes). P(Sports=Yes) / P(X)
= P(X | Sports=Yes). (3/5)
= P(A| Sports=Yes). P(Very | Sports=Yes). P(Close | Sports=Yes). P(Game | Sports=Yes). (3/5)
= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	Sports or Not Sports
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Laplace Smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

- In our example
- we add 1 to every probability, therefore the probability, such as **P(close | sports)**, will never be 0.

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

P(Sports=Yes) = 3/5
P(Sports=No) = 2/5
P(A | Sports=Yes) = 2/11
P(A | Sports=No) = 1/9

Sky
Sunny
Rainy
(2)

P(Sports=Yes | X) = P(X | Sports=Yes). P(Sports=Yes) / P(X)
= P(X | Sports=Yes). (3/5)
= P(A| Sports=Yes). P(Very | Sports=Yes). P(Close | Sports=Yes). P(Game | Sports=Yes). (3/5)
= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)
= (2+1/11+14) . (1+1/11+14) . (0+1/11+14) . (2+1/11+14) . (3/5)
= 0.00002765

(3+14/5+28)
=0.00002396

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	Sports or Not Sports
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Apply Laplace Smoothing

Word	P(word | Sports)	P(word | Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

P(new | Sports = Yes) > P(new | sports = No)

$$P(a|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times P(game|Not\ Sports) \times P(Not\ Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

The new instance/sentence will be categorized as "Sports"

Example 2: Multinomial model

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

N_{yes} (W = chinese) → no. of word counting 'chinese' in class "yes"

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)}.$$

N_{yes} (W=Chinese) = 5, N_{No} (W=Chinese) = 1,

unique words

|v| = 6 = {Chinese, Beijing, Shanghai, Macao, Tokyo Japan}

No of features (words) in Yes class = 8

No of features (words) in No class = 3

Source Credit : https://nlp.stanford.edu/IR-book

Example 2

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

P(Chinese | Yes) = 5/8

After Laplacian smoothing
P(Chinese | Yes) = (5+1)/(8+6)

Priors: $\hat{P}(c) = 3/4$ (Yes) and $\hat{P}(\overline{c}) = 1/4$ (No)

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)}$$

|V| → unique words

$$\hat{P}(\text{CHINESE}|c) = (5+1)/(8+6) = 6/14 = 3/7$$
$$\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) = (0+1)/(8+6) = 1/14$$
$$\hat{P}(\text{CHINESE}|\overline{c}) = (1+1)/(3+6) = 2/9$$
$$\hat{P}(\text{TOKYO}|\overline{c}) = \hat{P}(\text{JAPAN}|\overline{c}) = (1+1)/(3+6) = 2/9$$

Source Credit : https://nlp.stanford.edu/IR-book

Example 2

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C_k|\mathcal{D}) \propto P(C_k) \prod_{j=1}^{\text{len}(\mathcal{D})} P(u_j|C_k)$$

u-each word in test document

doc ID = 5

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$

$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

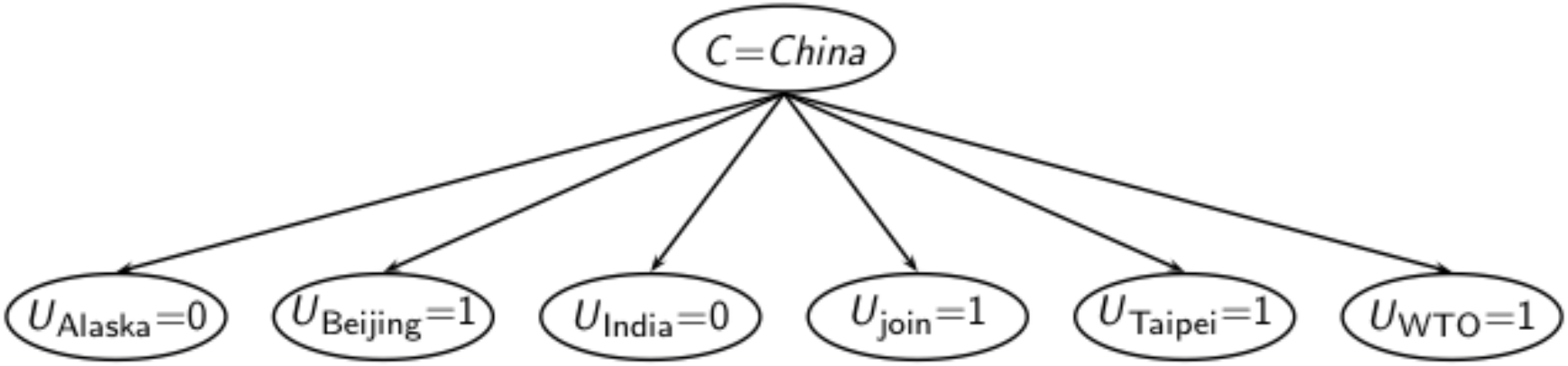
Posterior probability

Hence d_5 will be labelled as c = china.

since $P(c|d_5) > P(\bar{c}|d_5)$

Source Credit : https://nlp.stanford.edu/IR-book

Different Naive Bayes model: Bernoulli model



One feature X_w for each word in dictionary

X_w = true in document *d* if *w* appears in *d*

Source Credit : https://nlp.stanford.edu/IR-book

Example 3

Bernoulli Model

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{N_k},$$

Let $n_k(w_t)$ be the number of documents of class k in which w_t is observed; and let N_k be the total number of documents of that class.

N_{yes} (W=Chinese) = 3, N_{No} (W=Chinese) = 1,

No of features (documents) in Yes class : (N_{Yes}) = 3

No of features (documents) in No class :(N_{No}) = 1

Priors:

$$P(Yes) = \frac{3}{4}, P(No) = \frac{1}{4}$$

Source Credit : https://nlp.stanford.edu/IR-book

New Training set 0 1 1 0 0 1 Yes

Represent documents as binary (Bernoulli) features write all the unique words

docID	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	Class
1	1	1	0	0	0	0	Yes
2	1	0	1	0	0	0	Yes
3	1	0	0	1	0	0	Yes
4	1	0	0	0	1	1	No

Training set

Conditional probabilities (with Laplace smoothing):

$N_k = 4$

$n_k(\text{chinese}) = 3$

$n_k(\text{Tokyo}) = 1$

For Bernoulli NB:

$$P(w_i = 1|C_k) = \frac{n_k(w_i) + 1}{N_k + 2}$$

where:

- $n_k(w_i)$: number of documents in class C_k containing word w_i
- N_k: number of documents in class C_k
- Denominator has +2 because feature is binary (present or not)

docID=4 Tokyo Japan Chinese , class=NO

For class = No (1 document)

Word	Docs with word	$P(w|No) = (n+1)/(1+2)$
Chinese	1	2/3 ≈ 0.667
Beijing	0	1/3 ≈ 0.333
Shanghai	0	1/3 ≈ 0.333
Macao	0	1/3 ≈ 0.333
Tokyo	1	2/3 ≈ 0.667
Japan	1	2/3 ≈ 0.667

No. of documents in class 'Yes'

For class = Yes (3 documents)

Word	Docs with word	$P(w|Yes) = (n+1)/(3+2)$
Chinese	3	4/5 = 0.8
Beijing	1	2/5 = 0.4
Shanghai	1	2/5 = 0.4
Macao	1	2/5 = 0.4
Tokyo	0	1/5 = 0.2
Japan	0	1/5 = 0.2

Example 3

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$$
$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0+1)/(3+2) = 1/5$$
$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1+1)/(3+2) = 2/5$$
$$\hat{P}(\text{Chinese}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) = (0+1)/(1+2) = 1/3$$

Source Credit : https://nlp.stanford.edu/IR-book

Example 3

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

b = feature vector for the document D

b_t, = {0,1} => absence or presence of word w_t in the document

$$P(C_k|\mathbf{b}) \propto P(\mathbf{b}|C_k)\,P(C_k)$$

$$\propto P(C_k)\prod_{t=1}^{|V|}\left[b_t\,P(w_t|C_k)+(1-b_t)\,(1-P(w_t|C_k))\right].$$

$P(Yes|d_5)$

$$\begin{aligned}
\hat{P}(c|d_5) &\propto \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\\
&\quad\cdot(1-\hat{P}(\text{Beijing}|c))\cdot(1-\hat{P}(\text{Shanghai}|c))\cdot(1-\hat{P}(\text{Macao}|c))\\
&= 3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\\
&\approx 0.005\\
\hat{P}(\bar{c}|d_5) &\propto 1/4\cdot 2/3\cdot 2/3\cdot 2/3\cdot(1-1/3)\cdot(1-1/3)\cdot(1-1/3)\\
&\approx 0.022
\end{aligned}$$

$P(No|d_5)$ $P(\bar{c})\cdot P(\text{chinese}|\bar{c})\cdot P(\text{Japan}|\bar{c})\cdot P(\text{Tokyo}|\bar{c})\ldots$

Answer

P(No|d)>P(Yes|d)

Prediction = "No" (Not China)

Source Credit : https://nlp.stanford.edu/IR-book

Naïve Bayes classifier: Model Summary

- Naïve Bayes based on the independence assumption
  - Training is very easy and fast; just require considering each attribute in each class separately
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more…

Logistic Regression vs Naïve Bayes

Idea:

- Naïve Bayes allows computing P(Y|X) by learning P(Y) and P(X|Y)

- Why not learn P(Y|X) directly?

in high dimensional data

$P(Y|X_1, X_2 \ldots X_n)$

$P(Y|X_1, X_2)$

Logistic Regression vs Naïve Bayes

Aspect	Naïve Bayes	Logistic Regression
Model Type	Generative	Discriminative
Underlying Assumptions	Assumes features are independent of each other	No assumption about independence
Model Complexity	Relatively simple, computationally efficient.	More complex, allowing for modeling intricate relationships. Potential for overfitting with a higher number of features.
Handling Categorical Data	Well-suited for categorical data, effective for discrete features and text classification.	Versatile, can handle both numerical and categorical features. Requires encoding for categorical variables with multiple levels.
Interpretability	Highly interpretable due to simplicity and the assumption of feature independence.	Good interpretability through coefficients, indicating the strength and direction of feature relationships.
Robustness to Irrelevant Features	Can be robust to irrelevant features due to the assumption of independence.	May be sensitive to irrelevant features. Regularization techniques can help mitigate this sensitivity.
Data Size and Sparsity	Performs well with small datasets. Effective with sparse data, e.g., in text classification.	May require a larger dataset to avoid overfitting, especially with a high number of features. Can handle sparse data with careful regularization.
Applications	Suitable for Text classification, spam filtering, sentiment analysis	Suitable for customer churn prediction, credit scoring.

Logistic Regression and Gaussian Naïve Bayes Classifier

- Interestingly, the parametric form of P(Y|X) used by Logistic Regression is precisely the form implied by the assumptions of a Gaussian Naive Bayes classifier.

- Therefore, we can view Logistic Regression as a closely related alternative to GNB, though the two can produce different results in many cases

- Derivation given in CS-5 slides in self learning section and also attached PDF of new chapter from Tom Mitchell book

Features of Bayesian learning

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

- Flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

$P(y|x)$ → depends on $P(y)$ $P(x|y)$

References

- https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn07-notes-nup.pdf
- https://cs229.stanford.edu/summer2019/cs229-notes2.pdf
- Tom Mitchell – Chapter 6
- https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

Logistic Regression & Naïve Bayes

Self Read

(Proof : Mathematical Relation between Logistic Regression and Naïve Bayes)

Where does the **form** come from?

MLE Revision

$$\begin{aligned}\theta_{\text{MLE}} &= \arg\max_{\theta} L(y_1, \ldots, y_n|\theta) \\ &= \arg\max_{\theta} \log L(y_1, \ldots, y_n|\theta) \ ; \text{maximising } f \equiv \text{maximising } \log f \\ &= \arg\max_{\theta} \log \prod_{i=1}^{n} L(y_i|\theta) \\ &= \arg\max_{\theta} \sum \log L(y_i|\theta) \ldots\ldots (1.5)\end{aligned}$$

Generative Classifier from bayes Rule

$$p(C_k|x) = \frac{p(x|C_k)\, p(C_k)}{p(x)}$$

Slide credit: Tom Mitchell

Where does the **form** come from?

Applying Bayes rule

$$p(Y = c_k|X_1, X_2, \ldots, X_d) = \frac{p(X_1, X_2 \ldots, X_d|Y = c_k)\, p(Y = c_k)}{\sum_{k=1}^{K} p(X_1, \ldots, X_d|Y = c_k)\, p(Y = c_k)} \ldots$$

If we use the assumption of conditional independence

$$p(Y = c_k|X_1, X_2, \ldots, X_d) = \frac{p(Y = c_k) \prod_{i=1}^{d} p(X_i|Y = c_k)}{\sum_{k=1}^{K} p(Y = c_k) \prod p(X_i|Y = c_k)} \ldots\ldots$$

$$\hat{Y} = \arg\max_{y_k} p(Y = c_k|\tilde{X}_1, \ldots, \tilde{X}_d)$$

$$= \arg\max_{y_k} \frac{p(\tilde{X}_1, \ldots, \tilde{X}_d|Y = c_k)\, p(Y = y_k)}{\sum_{k=1}^{K} p(\tilde{X}_1, \ldots, \tilde{X}_d|Y = y_k)\, p(Y = c_k)} .$$

$$= \arg\max_{c_k} p(Y = c_k) \prod p(\tilde{X}_i|Y = c_k)$$

Logistic Regression –Bayesian Analysis

Applying Bayes rule

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)}$$

$$P(Y=1|X) = \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}$$

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{P(Y=0)}{P(Y=1)}+\ln\frac{P(X|Y=0)}{P(X|Y=1)}\right)}$$

Logistic Regression –Bayesian Analysis

By independence assumption:

$P(Y=1)=\pi$ and $P(Y=0)=1-\pi$
by modelling P(Y) as Bernoulli

$$\frac{P(X|Y=0)}{P(X|Y=1)} = \prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \ln\prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

Logistic Regression –Bayesian Analysis

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right))}$$

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

$$w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

$$\begin{aligned}
\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} &= \sum_i \ln\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)} \\
&= \sum_i \ln\exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)
\end{aligned}$$

Some Additional References

- T1 book by Tom Mitchell – Chapter-6
- https://web.stanford.edu/class/archive/cs/cs109
- https://cs229.stanford.edu/lectures-spring2022/main_notes.pdf
- https://www.cs.cmu.edu/~ninamf/courses/601sp15/lectures.shtml
- https://nlp.stanford.edu/IR-book - Chapter 13

Thank you !

Required Reading for completed session :

T1 - Chapter # 8 (Tom M. Mitchell, Machine Learning)

T1 - Chapter # 6 (Tom M. Mitchell, Machine Learning)

R1 – Chapter # 3 (Christopher M. Bhisop, Pattern Recognition & Machine Learning)

& Refresh your MFDS & ISM parallel course basics

Next Session Plan :

Module 9 : Ensemble Learning